Homework 2 Solutions

Problem 2.2

Let us assume that there exist scalars $c_0 \geq c > 0$ such that for all $k \in \mathbb{N}$

$$\nabla f(\mathbf{w}_k)^T \mathbb{E}_{\zeta_k}\left[g(\mathbf{w}_k;\zeta_k)\right] \geq c\|\nabla f(\mathbf{w}_k)\|_2^2, \tag{1a}$$

$$\|\mathbb{E}_{\zeta_k}\left[g(\mathbf{w}_k;\zeta_k)\right]\|_2 \leq c_0\|\nabla f(\mathbf{w}_k)\|_2. \tag{1b}$$

Furthermore, let us assume that there exist scalars $M \geq 0$ and $M_V \geq 0$ such that for all $k \in \mathbb{N}$

$$\mathrm{Var}_{\zeta_k}\left[g(\mathbf{w}_k;\zeta_k)\right] \leq M + M_V\|\nabla f(\mathbf{w}_k)\|_2^2. \tag{2}$$

For the convergence proof of SGD with an L-smooth convex objective function (see slides), prove that

$$\mathbb{E}_{\zeta_k}\left[\|g(\mathbf{w}_k;\zeta_k)\|_2^2\right] \leq \alpha + \beta\|\nabla f(\mathbf{w}_k)\|_2^2.$$

proof: The variance of $g(\omega_k, \zeta_k)$ is

$$\mathrm{Var}_{\zeta_k}\left[g(\omega_k, \zeta_k)\right] = E_{\zeta_k}\left[\|g(\omega_k, \zeta_k)\|_2^2\right] - \left\|E_{\zeta_k}\left[g(\omega_k, \zeta_k)\right]\right\|_2^2$$

$$\Rightarrow E_{\zeta_k}\left[\|g(\omega_k, \zeta_k)\|_2^2\right] - \left\|E_{\zeta_k}\left[g(\omega_k, \zeta_k)\right]\right\|_2^2 \overset{(2)}{\leq} M + M_V\|\nabla f(\omega_k)\|_2^2$$

$$\Rightarrow E_{\zeta_k}\left[\|g(\omega_k, \zeta_k)\|_2^2\right] \leq \left\|E_{\zeta_k}\left[g(\omega_k, \zeta_k)\right]\right\|_2^2 + M + M_V\|\nabla f(\omega_k)\|_2^2$$

$$\overset{(1b)}{\leq} c_0^2\|\nabla f(\omega_k)\|_2^2 + M + M_V\|\nabla f(\omega_k)\|_2^2$$

$$= M + \beta\|\nabla f(\omega_k)\|_2^2$$

where $\beta \triangleq c_0^2 + M_V \geq \mu^2 > 0$ $\square$

(M could be α)

3) Show that $\lim_{k\to\infty} \mathbb{E}\left[\sum_{i=1}^{K} \alpha_k \|\nabla \cancel{f_i} f(w_k)\|_2^2\right] < \infty$.

Solution:

One assumption is than $\alpha_k, \forall k$ are square summable, meaning $\sum_{k=1}^{K} \alpha_k \xrightarrow{k\to\infty} < \infty$.

As $K\to\infty$, we can say that $\{\alpha_k\} \to 0$, and $\alpha_k L M_G \leq \mu$ for all $k \in [K]$.

~~Another assum~~

For a single iteration, we can say that

$$\mathbb{E}_{\xi_k}[f(w_{k+1})] - f(w_k) \leq \alpha_k \nabla f(w_k)^T \underbrace{\mathbb{E}_{\xi_k}[g(w_k, \xi_k)]}_{\Delta} + \frac{1}{2}\alpha_k^2 L \underbrace{\mathbb{E}_{\xi_k}[\|g(w_k, \xi_k)\|_2^2]}_{*}$$

Here, we will use 2 properties (results) from the assumption that variance of the g is limited.

$$*: \mathbb{E}_{\xi_k}[\|g(w_k, \xi_k)\|_2^2] \leq M + M_G \|\nabla f(w_k)\|_2^2 \quad \text{with } M_G = M_V + \mu_G^2 \geq \mu^2 > 0$$

$$\Delta: \nabla f(w_k)^T \mathbb{E}_{\xi_k}[g(w_k, \xi_k)] \geq \mu \|\nabla f(w_k)\|_2^2$$

By replacing these terms, we obtain

$$\mathbb{E}_{\xi_k}[f(w_{k+1})] - f(w_k) \leq -\left(\mu - \frac{1}{2}\alpha_k L M_G\right)\alpha_k \|\nabla f(w_k)\|_2^2 + \frac{1}{2}\alpha_k^2 L M$$

Now we take expect. of both sides

$$\mathbb{E}[f(w_{k+1})] - \mathbb{E}[f(w_k)] \leq -\left(\mu - \frac{1}{2}\alpha_k L M_G\right)\alpha_k \mathbb{E}[\|\nabla f(w_k)\|_2^2] + \frac{1}{2}\alpha_k^2 L M$$

By using square summ. prop. $\alpha_k L M_G \leq \mu$

$$\leq -\frac{1}{2}\mu \alpha_k \mathbb{E}[\|\nabla f(w_k)\|_2^2] + \frac{1}{2}\alpha_k^2 L M.$$

Now we will iterate over K terms, and sum all iterations:

$$f^* - \mathbb{E}[f(w_K)] \leq \mathbb{E}[f(w_{K+1})] - \mathbb{E}[f(w_1)] \leq -\frac{1}{2}\mu \sum_{k=1}^{K} \alpha_k \mathbb{E}[\|\nabla f(w_k)\|_2^2] + \frac{1}{2} L M \sum_{k=1}^{K} \alpha_k^2$$

$$\sum_{k=1}^{K} \alpha_k \mathbb{E}[\|\nabla f(w_k)\|_2^2] \leq \underbrace{\frac{2(\mathbb{E}(f(w_1)) - f^*)}{\mu}}_{<\infty} + \underbrace{\frac{LM}{\mu}\sum_{k=1}^{K} \alpha_k^2}_{<\infty}$$

Due to square summability

$\square$